ChinaXiv合作期刊

知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

文章关键词和标题分词多样性及其绘制知识图谱的比较

李继红 徐桂珍 江珊 王洪江

安徽省农业科学院农业经济与信息研究所 合肥 230031

摘要：[目的/意义]基于文章的关键词和标题分词，分析关键词和标题分词的多样性以及基于二者绘制的知识图谱的差异。[方法/过程]爬取中国知网 2010 – 2019 年研究主题为"学术不端"的相关论文，采用多样性指数定量分析文章关键词和标题分词的特征，并通过 CiteSpace 软件定性比较基于关键词和标题分词所绘制知识图谱的架构。[结果/结论]关键词的丰富度（S）、多样性（H'）和均匀度指数(E_H)均异于标题分词，且两个单元的相似性较弱，表明文章关键词和标题分词是两个不同的单元；基于此绘制的知识图谱虽有差异，但二者均能从各自的角度展示"学术不端"领域的研究主题。

关键词：学术不端 关键词 标题 中文分词 多样性 知识图谱

分类号：G250

DOI：10.13266/j.issn.2095-5472.2021.005

引用格式：李继红，徐桂珍，江珊，等．文章关键词和标题分词多样性及其绘制知识图谱的比较 [J/OL]. 知识管理论坛，2021, 6(1): 46-55[引用日期]. http://www.kmf.ac.cn/p/239/.

知识图谱是通过将应用数学、图形学、信息科学等学科的理论、方法与计量学引文分析、共现分析等方法结合，并利用可视化的图谱形象地展示学科的核心结构、发展历史、前沿领域以及整体知识架构达到多学科融合的现代理论。它可以把复杂的知识领域通过数据挖掘、信息处理、知识计量和图形绘制显示出来，揭示知识领域的动态发展规律，为学科研究提供有价值的参考[1]。目前，知识图谱的绘制工具有多种类型，主要包括 CiteSpace、HistCite、Sci2 Tools、Leydesdorff、Ucinet、Pajek、VOSviewer 等[2]。其中，CiteSpace 是应用最广泛、功能最强大的信息可视化软件，可通过选择节点类型进行相应的共被引网络、共现网络或合作网络的分析，进而形成可视化、序列化的知识图谱[3-5]。

在所发表的 CiteSpace 相关论文中，对关键词进行共现分析的占了较大比例。关键词是为了文献标引工作，从报告、论文中选取出来以

作者简介：李继红（ORCID:0000-0002-1863-0890），助理研究员，博士；徐桂珍（ORCID: 0000-0003-1304-8831），研究员，硕士；江珊（ORCID:0000-0001-5127-0367），副研究员，硕士；王洪江（ORCID: 0000-0002-2456-5466），研究员，硕士，通讯作者，E-mail: 2398606371@qq.com。

收稿日期：2020-12-21 发表日期：2021-02-24 本文责任编辑：易飞

知识管理论坛，2021（1）:46-55
DOI: 10.13266/j.issn.2095-5472.2021.005

表示全文主题内容信息款目的单词或术语[6]。从文献库存储信息的特点形式来说，关键词是摘要的“摘要”，高度概括了论文主题，集中表达了论文内容的核心和精髓。对论文的关键词进行共现分析，可以探讨研究领域的热点、趋势以及知识结构等。而对于没有关键词的数据源（论文标题、基金项目、网络舆情、影评）进行分析时，主要采用中文分词的方法[7-9]。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。论文标题是标明文章内容的简短语句，是文章最重要的部分。对论文标题进行分词，可以得到涵盖文章内容和主旨的词语。

论文关键词和标题分词都能揭示论文的主题内容，那么对于同一篇文章，关键词和标题分词有何关联，基于二者绘制的知识图谱又存在什么差异？本研究以中国知网（CSSCI 来源期刊）为统计源，爬取以“学术不端”为主题的研究文献，并采用多样性指数定量分析关键词和标题分词的特征，依托 CiteSpace 软件定性分析基于关键词和标题分词所绘制知识图谱的架构差异。该研究不仅可以为相关研究提供一定的理论基础，还有着积极的应用意义。

1 数据来源及分析方法

1.1 数据来源

笔者以中国知网（CSSCI 来源期刊）为数据统计源，检索研究主题为“学术不端”的学术论文。检索式为：“主题 = 学术不端 or 学术造假 or 学术道德 or 科研诚信”；论文发表时间为 2010 － 2019 年。检索时间为 2019 年 10 月 24 日，剔除通知、征文、声明等，清洗后共得到 757 条学术论文，再爬取、汇总题录信息，题录信息主要包括论文标题、作者、关键词等字段。

1.2 分析方法

1.2.1 标题分词

对论文标题进行分词处理应遵循如下原则：①应选择能明确表达主题概念的词语；②筛除无意义的词语；③对名称内含义相同但是表述不同的词汇进行合并。例如，学术不端、学术腐败等词规范化处理后统一为“学术不端行为”，AMLC、学术不端检测系统等统一为“学术不端文献检测系统”。

基于以上分词原则，采用武汉大学研发的 ROSTCM6 软件对项目名称进行分词。直接获得的关键词的粒度比较粗糙，聚类效果不理想，可通过人工标注的方式补充添加用户词典，使其达到研究要求。再采用数据清洗器对分词后的词语进行清洗、合并。

1.2.2 关键词和标题分词多样性分析

多样性一直常见于生态学名词，常用的多样性指数主要包括丰富度指数（Richness index，S）、Shannon-Wiener 多样性指数（Shannon-Wiener diversity index，H'）、Pielou 均匀度指数（Pielou evenness index，E_H）等，可用于判断群落或生态系统的多样性、复杂性[10-12]。本研究中，借用上述 3 个指数来描述关键词和标题分词的多样性。

$$H' = -\sum_{i=1}^{S} p_i \ln p_i = -\sum_{i=1}^{S} (n_i / N)\ln(n_i / N)$$ 公式（1）

$$E_H = H' / H'_{\max} = H' / \ln S$$ 公式（2）

式中，S 为关键词或标题分词的词语数量；P_i 为关键词或标题分词词语 i 的相对丰度，代表某一词语的数量在所有词语总量中所占的比率，即 $P_i = n_i/N$，n_i 是关键词或标题分词词语 i 的数量，N 是所有关键词或标题分词的数量。

为了对关键词和标题分词两个单元的相似性进行研究，笔者借用 Sørensen 指数（C_S）和 Jaccard 指数（C_J）对二者进行分析。Sørensen 指数和 Jaccard 指数是生态学中用于反映群落间物种组成相似性的指数[13]，这里用来反映单元间词语的相似性。

$$C_S=2c/(a+b)$$ 公式（3）

$$C_J=c/(a+b-c)$$ 公式（4）

式中，c 为关键词和标题分词两个单元的共有词语数；a 和 b 分别为关键词和标题分词的词语数。

1.2.3 标题分词格式转化

众所周知，CiteSpace 软件只能分析特定数据库中的文献，还不能直接用于其他数据库。笔者采用格式转化软件对非特定数据库中的数据进行转化处理，使之成为 CiteSpace 软件能够识别的数据，从而进行相关的分析。

1.2.4 关键词和标题分词共现分析

关键词或标题分词共现分析就是对数据集中关键词或标题分词集合进行分析，通过对关键词或标题分词的可视化分析可以确定研究领域的学科结构、研究热点等。笔者分别采用“学术不端”研究文献的关键词和标题分词集合为分析单元，依托 CiteSpace 绘制主题聚类图，从而比较该领域的知识架构。

可视化分析的参数设置如下：时间跨度设置为 2010 – 2019 年，时间切片（Time Slicing）为 1 年；节点类型（Node Types）确定为 keyword；节点强度（Links）默认 Cosine 与 Within Slices 选项；选择阈值（Selection Criteria）选取 Top N per slice=50；网络裁剪功能区（Pruning）默认不进行剪裁，最终生成关键词和标题分词共现知识图谱。

2 关键词和标题分词的多样性比较

2.1 关键词和标题分词的词语组成

笔者对“学术不端”相关文献的关键词和标题分词进行统计分析，分别得到 3 131 个关键词和 3 094 个标题分词，把各个单元的相同项进行整理，最终获得 1 541 个关键词词语和 1 432 个标题分词词语。

关键词和标题分词出现的频次以及该词频下词语的数量见表 1。从表 1 可以看出，出现频次最高（344 次）的关键词是学术不端行为，然后依次是科研诚信（95 次）和学术道德（77 次），出现频次最少的为 1 次。随着词频的下降，该词频下关键词的数量呈上升趋势，例如，学术不端行为、科研诚信、学术道德、研究生的词频较高，该词频下的关键词数量较少（1 个）；而词频为 3 次以下的关键词则较多，词频为 2 的关键词为 147 个，词频为 1 的有 1 254 个，占比高达 40%。出现频次最高的标题分词也是学术不端行为，达到 268 次；排名第 2 和第 3 的分别是研究生和高校，出现频次分别为 107 和 84 次；出现最少的词频也是 1 次，出现频次为 1 的词语共有 1 094 个，占所有标题分词的 35.36%。

对于“学术不端”的研究，论文关键词比标题分词多 37 个；所整理的词语，前者比后者多 109 个（7.61%），说明论文自带的关键词比标题分词后的词语要丰富。但出现频次较高的一些词语还是比较一致的，比如学术不端行为、研究生、高校、科研诚信、科技期刊等。说明不管用标题分词还是论文关键词，最核心的词语是不变的，而且在这两种方法中，随着词频的下降，该词频下的词语数量均呈现上升趋势。

2.2 关键词和标题分词的多样性

基于文章关键词和标题分词的词频以及该词频下词语的数量，本研究对这两个单元（关键词和标题分词）的词语多样性进行了分析。用 CiteSpace 可视化软件绘制知识图谱时，词语的出现频次设定阈值为 ≥2 次，因此除了对两个单元内所有词语进行统计外，还对出现频次 ≥2 词语的多样性进行了分析。

研究主要采用丰富度指数（S）、Shannon-Wiener 多样性指数（H'）、均匀度指数（E_H）、Sørensen 指数（C_S）和 Jaccard 指数（C_J），从单元内、单元间两个层面对文章的关键词和标题分词进行多样性的分析，以探索表达相同主题而来源不同的词语在数量、丰度、分布情况等方面的差异以及二者的相似性。其中，丰富度指数、Shannon-Wiener 多样性指数、均匀度指数属于 α 多样性指数，主要用于研究单元内词语的结构多样性。丰富度指数（S）的大小反映了词语数量的多少；Shannon-Wiener 多样性指数（H'）是基于词语数量来反映单元内词语的多样性，H' 值越大，表示单元所含的信息量就越大，词语的多样性就越高；均匀度指数（E_H）可反映单元内词语的均匀度，E_H 数值越

知识管理论坛，2021（1）:46-55
DOI: 10.13266/j.issn.2095-5472.2021.005

高，表明各个词语的数量越接近；Sørensen 指数和 Jaccard 指数属于 β 多样性指数，主要用于分析研究单元间词语的相似性，数值越大，说明两个单元越相似，一致性越高。

表 1 关键词和标题分词的统计性描述

关键词	词频（次）	词语数量（个）	标题分词	词频（次）	词语数量（个）
学术不端行为	344	1	学术不端行为	268	1
科研诚信	95	1	研究生	107	1
学术道德	77	1	高校	84	1
研究生	67	1	科研诚信	48	1
学术规范	49	1	学术道德	46	1
科技期刊	46	1	科技期刊	43	1
学术期刊	44	1	治理	34	1
学术不端文献检测系统	42	1	对策、学术不端文献检测系统	29	2
高校	38	1	启示	28	1
应对策略	31	1	科技论文、美国	24	2
编辑	23	1	学术期刊、期刊编辑、学术规范等	23	4
学术不端文献检测	19	1	大学生、高校教师、实证分析	18	4
防范策略	16	1	比较、学术诚信、学术道德失范	17	3
美国	15	1	现状	16	1
研究生教育、高校教师、大学生	14	3	实践	15	1
学术失范	13	1	制度	14	2
学位论文、学术腐败	12	2	论文撤销、学术道德教育	13	2
治理、文字复制比、期刊编辑等	11	8	学位论文	12	1
影响因素、一稿多投、学术论文	9	4	作用	11	1
医学期刊、学术评价、学术共同体等	8	8	检测、特点	10	2
学术责任、学术生态、学术伦理等	7	7	成因、规范、科技期刊编辑	9	7
制度建设、审稿专家、数据库等	6	10	科研人员、路径、学术论文	8	3
自律、著作权、同行评议等	5	13	策略、创新、期刊	7	8
重复率、治理体系、知识产权等	4	25	撤销论文、高校图书馆、医学论文等	6	5
学术制度、学术环境、实证分析等	3	44	诚信、存在问题、管理等	5	15
作者、职称评定、职业伦理等	2	147	伦理学、科研管理、学术行为等	4	24
作者资格、作者信息、自我剽窃等	1	1 254	案例分析、编辑部、博士生等	3	61
			CNKI、参考文献、大数据等	2	181
			Scopus、SWOT分析、科技查新等	1	1 094

关键词和标题分词的多样性见表 2。从表 2 中可看出，对于所有词频的关键词，其丰富度指数为 1 541，Shannon-Wiener 指数为 6.25，均匀度指数为 0.85；对于所有词频的标题分词，其丰富度指数为 1 432，Shannon-Wiener 指数为 6.26，均匀度指数为 0.96。关键词的丰富度大于标题分词的，二者的 Shannon-Wiener 指数较接近，关键词的均匀度指数小于标题分词的，这是由于后者各个词语的数量比前者更接近，分布更均匀。因为两个单元内词频为 1 的词语数量较多，所以词频 ≥2 的关键词和标题分词的丰富度均大幅下降，关键词的丰富度下降了 81.38%，标题分词的丰富度减少了 76.40%，前者的丰富度（287）和多样性（4.54）均小于后者的丰富度（338）和多样性（4.86），但二者的均匀度相近。Sørensen 指数（C_S）和 Jaccard 指数（C_J）是用来衡量两个单元相似度的指标。在本研究中，所有频次的关键词和标题分词两个单元的相似性较低，C_S 和 C_J 的数值分别为 0.39 和 0.24；对于频次 ≥2 的词语，两个单元的相似性比前者稍高，C_S 和 C_J 分别提高了 17.95% 和 25%，但也是弱相关。

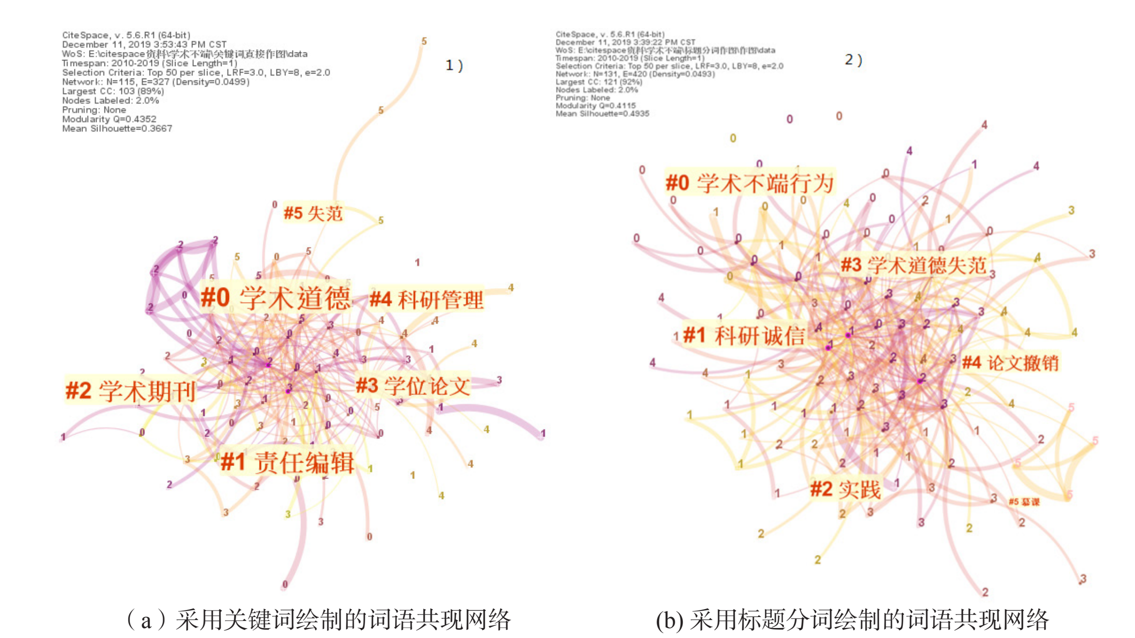
表 2 关键词和标题分词的多样性

项目	S	H'	E_H	C_S	C_J
所有关键词	1 541	6.25	0.85	0.39	0.24
所有标题分词	1 432	6.26	0.96		
≥2次关键词	287	4.54	0.80	0.46	0.30
≥2次标题分词	338	4.86	0.83		

3 利用关键词和标题分词绘制知识图谱

3.1 图谱参数比较

基于 2010 － 2019 年间发表论文的关键词和标题分词，按照统一的参数设置，可以生成如图 1 所示的共现网络。需要说明的是，软件选取每一年中被引次数排名居前 50 位的词语，且每个词语至少出现 2 次。在图 1 中，数字代表聚类主题的 ID 号；每个节点代表一个关键词，节点的大小表示关键词的出现频次。



（a）采用关键词绘制的词语共现网络　　(b) 采用标题分词绘制的词语共现网络

图 1 采用关键词和标题分词绘制的知识图谱

知识管理论坛，2021（1）:46-55
DOI: 10.13266/j.issn.2095-5472.2021.005

知识图谱左上角的数据是该图谱的参数，主要包括使用软件信息、运行时间、数据存放位置、时间切片（Timespan）、选择标准（Selection Criteria）、生成网络（Network）、裁剪方法（Pruning）、模块度（Modularity）以及平均轮廓值（Mean Silhouette）等。其中，软件信息、运行时间、数据存放位置、时间切片、选择参数属于软件设置参数，生成网络、最大的分支、裁剪方法、模块度以及平均轮廓值属于网络生成参数，可用于解读图谱的信息。表 3 为分别采用关键词和标题分词所绘制图谱的参数，包括 N、E、Density、Modularity、Silhouette、Largest CC。N 表示网络节点数量；E 表示连线数量；Density 表示网络的密度；Modularity 表示网络的模块度，值越大表示网络的聚类结果越好；Mean Silhouette 表示聚类平均轮廓值，Silhouette 值是用来衡量网络同质性的指标，越接近 1，网络的同质性越高。

表 3 基于关键词和标题分词所绘制知识图谱的参数

词语来源	N	E	Density	Modularity	Silhouette
关键词	115	327	0.049 9	0.435 2	0.366 7
标题分词	131	420	0.049 3	0.411 5	0.493 5

从表 3 可以看出，基于关键词绘制的图谱，其网络节点数有 115 个，连线数有 327 条，网络密度为 0.049 9；基于标题分词绘制的图谱，网络节点数和连线数分别为 131 个和 420 条，比前者分别增加了 13.91% 和 28.44%，网络密度为 0.049 3，和前者相近。Modularity 和 Mean Silhouette 是反映图谱整体框架特征的重要参数。在采用关键词绘制的图谱中，Modularity 值和 Silhouette 值分别为 0.435 2 和 0.366 7。在采用标题分词绘制的图谱中，Modularity 值为 0.411 5，比前者低 5.7%；Silhouette 值为 0.493 5，比前者高 34.58%，说明前者所有集群的同质化程度低于后者。二者的 Modularity 值均大于 0.3，一般认为聚类模块值 >0.3 意味着聚类结构显著，说明无论是采用关键词还是标题分词绘制的图谱，其结构均符合聚类要求。

3.2 图谱词语比较

在图谱中，词语的频次高低能够反映出该领域研究的总体状况，每一个词语对应图谱上的一个节点。即采用关键词和标题分词绘制共现图谱的词语分别为 115 个和 131 个。

表 4 为基于关键词和标题分词绘制图谱中词频 ≥30 的词语信息，包括词语、词语出现的词频以及其中介中心性。中介中心性是测定节点在网络中重要性的一个指标，是一个用以量化点在网络中地位重要性的图论概念 [2]。词语的中介中心度越大，说明其在图谱中的重要性越大。在采用关键词绘制的图谱中，词频 ≥30 的词语有 9 个，词频加起来共 782 次。其中，词频最高的学术不端行为，共出现 340 次，其次为科研诚信（92 次）、学术道德（76 次）、研究生（67 次），其中介中心性分别为 0.38、0.23、0.2、0.32。虽然词语的频次排序与中介中心度并非一一对应，但在整体上是基本一致的。在采用标题分词绘制的图谱中，词频 ≥30 的词语有 7 个，词频 639 次，词语和词频量均小于前者。但主要的词语和前者的相近，都包括了学术不端行为、科研诚信、学术道德、研究生、科技期刊以及高校等，词频最高的词语也是学术不端行为（262 次），其中介中心性最高（0.39）。

3.3 图谱聚类比较

聚类分析法是一种探索性数据挖掘分析方法，可用于识别和分析特定研究领域中显著术语和背景的分类，利用一系列的算法将收集到的数据转换成几个结构化的集群，从而发现知识领域的主题分布和组织结构 [14]。

表 4 基于关键词和标题分词绘制图谱中词频≥ 30 的词语（节点）信息

序号	词频	关键词	中介中心性	序号	词频	标题分词	中介中心性
1	340	学术不端行为	0.38	1	262	学术不端行为	0.39
2	92	科研诚信	0.23	2	107	研究生	0.37
3	76	学术道德	0.20	3	84	高校	0.29
4	67	研究生	0.32	4	64	科研诚信	0.20
5	48	学术规范	0.09	5	45	学术道德	0.16
6	45	科技期刊	0.18	6	43	科技期刊	0.07
7	43	学术期刊	0.23	7	34	治理	0.10
8	37	高校	0.13				
9	34	学术不端文献检测系统	0.06				

从图 1 可以看出，采用关键词绘制的图谱，共聚合成 6 个主题，集群从大到小依次为 #0 学术道德、#1 责任编辑、#2 学术期刊、#3 学位论文、#4 科研管理和 #5 失范。在采用标题分词绘制的图谱中，也聚合成 6 个主题，从大到小依次为 #0 学术不端行为、#1 科研诚信、#2 实践、#3 学术道德失范、#4 论文撤销和 #5 慕课。两个图谱中每个集群的信息见表 5，涵盖了各个集群包含的节点以及该群的轮廓值（Silhouette）。从表 5 可以看出，在采用关键词绘制的图谱中，各个集群的轮廓值都较高；在采用标题分词绘制的图谱中，除了 #1 的轮廓值稍低（0.375），其他集群的同质化程度都很高，再结合图 1 中的参数情况，可以得出，两幅图在聚类方面是理想的。

表 5 基于关键词和标题分词绘制图谱中各集群信息

关键词图谱群号	大小	Silhouette	标题分词图谱群号	大小	Silhouette
0	23	0.623	0	28	0.868
1	20	0.787	1	28	0.375
2	19	0.774	2	22	0.749
3	15	0.806	3	21	0.816
4	15	0.715	4	18	0.645
5	11	0.796	5	4	0.963

根据聚类主题的语义结构和研究主题的相关性，分别将二者的集群进行整合。采用关键词绘制的图谱可整合为三大知识域，分别是学术不端的行为和该方向研究的两大主要领域（期刊和高校）。第一个知识域包括 #0 学术不端行为和 #5 失范，研究主题涵盖学术不端的具体表现；第二个知识域包括 #1 责任编辑和 #2 学术期刊，反映了期刊是该领域的研究重点；第三个知识域涵盖 #3 学位论文和 #4 科研管理，体现了高校是学术不端研究的另一重要领域。采用标题分词绘制的图谱中的聚类也可整合为三大知识域：第一个知识域包括 #0 学术不端行为、#1 科研诚信和 #3 学术道德失范，说明学术不端研究的问题主要集中在学术不端行为、科研诚信、学术道德失范等方面；第二个知识域涵盖 #2 实践和 #5 慕课，主要体现了学术不端的防范，这一问题又可以分为素养教育培训和体系构建两个维度；第三个知识域即 #4 论文撤销，主要

知识管理论坛，2021（1）:46-55
DOI: 10.13266/j.issn.2095-5472.2021.005

研究学术不端的后果以及撤销论文带来的影响等。

采用关键词和标题分词绘制的图谱，其聚类结构既有相同，也存在一定的差异，这与词语的来源相关。来源为关键词的一部分属于表达核心主题因素的词语，可表达论文主题的关键性因素；一部分属于非核心主题因素的词语[15]，包括对核心主题因素起限定修饰作用的概念、核心主题因素的具体研究内容、研究过程中所应用的新方法及改进的常规方法、对核心主题因素起限定作用的时间和空间因素等，这两类词语共同概括了文章的主题、表达了内容的核心。来源为标题分词的是对文章标题进行分词而产生的，标题是文章精要内容的提炼、概括与浓缩，切分后大多数属于表达核心主题因素的词语，而非核心主题因素的词语较少，可能会缺少某些非核心因素、补充性的词语，从而在一定程度上有别于文章的关键词。采用关键词和标题分词绘制的图谱均可清晰、客观地展现学术不端研究领域的研究主题，但由于词语来源不同、性质不同，图谱所表达的侧重点亦不同。采用关键词绘制的图谱侧重于体现学术不端研究的问题，采用标题分词绘制的图谱更倾向于学术不端研究的方式方法。

4 讨论与结论

（1）CiteSpace 软件的应用拓展。如何从海量的文献信息中快速厘清从事领域的研究架构，找到最重要、最关键的有效信息，了解其过去、现在及趋势，是科学研究中面临的难题。知识图谱的出现为解决上述难题提供了有益的科学探索途径。信息可视化软件 CiteSpace 是一款功能强大的工具，所绘制的图谱具有“一图展春秋，一览无余；一图胜万言，一目了然”的特点[2]，从其问世便得到广泛的应用。目前，CiteSpace 软件只能用于分析特定数据库中的文献信息，包 括 WoS、Scopus、ADS、arXiv、CNKI、CSSCI、NSF、CSCD、Derwent 专利数据库等，而对于上述数据库以外的数据信息，还不能直接进行分析。笔者采用格式转化软件对非指定数据库中的数据进行格式的转化处理，使其成为 CiteSpace 软件能够识别分析的数据。研究结果显示，该方法科学有效，拓展了 CiteSpace 软件的应用数据源，可以为非 CiteSpace 指定数据库数据的可视化分析提供参考。

（2）关键词与标题分词的多样性。语言作为逻辑思维和推理工具，其基本要素是语词[16]。笔者以学术文献中的关键词和标题分词作为概念演化基础，尝试采用丰富度指数（S）、Shannon-Wiener 多样性指数（H'）、均匀度指数（E_H）、Sørensen 指数（C_S）和 Jaccard 指数（C_J）等比较两种词语的多样性。

对于“学术不端”的研究，论文关键词比标题分词的词语多 109 个，但出现频次较高的一些词语还是一致的。说明不管用标题分词还是论文关键词，其最核心的词语是相同的。对于所有词语而言，关键词的丰富度大于标题分词，多样性指数二者较接近。因为标题分词各个词语的数量比前者分布更均匀，所以其均匀度稍高。由于去除了词频等于 1 的大量词语，对于词频 ≥2 的关键词和标题分词，其词语的丰富度比所有词语时均大幅下降。关键词的丰富度和多样性均小于后者，但二者的均匀度较相近。在本研究中，不管是所有频次的关键词和标题分词还是词频大于 2 的词语，两个单元的相似性都较弱，说明二者是差异较大的两个单元，这为后续知识图谱的绘制提供了支撑。

（3）关键词和标题分词的共现网络。关键词是为了便于文献索引、文献标引和检索全文，并从论文中选取出来表示全文主题内容的词或词组。在对常规数据库中的数据进行分析时，CiteSpace 软件会自动提取文献的关键词，这些关键词既包括表达核心主题因素的词语，又包括非核心主题因素的词语。在本研究中，还通过对论文标题进行分词来获取词语，所获取的词语大多数属于表达核心主题因素的词语，而非核心主题因素的词语较少。关键词和标题分词都包含了表达核心主题因素的和非核心主题

因素的词语，但词语的数量和内容还是存在差异的，因此基于关键词和标题分词绘制的图谱，既有相同，也存在一定的差异。相同的是，两种方式绘制的知识图谱均能清晰、客观地展现“学术不端”研究领域的相关主题。不同的是，虽然采用同样的参数设置，但两种方式从各自的维度出发，揭示了不同的“学术不端”领域研究主题：采用关键词绘制的图谱更侧重于体现学术不端研究的问题，采用标题分词绘制的图谱则更倾向于学术不端研究的方式方法。

参考文献：

[1] 杨思洛，韩瑞珍. 国外知识图谱绘制的方法与工具分析 [J]. 图书情报知识，2012(6): 101-109.

[2] 李杰，陈超美. Citespace: 科技文本挖掘及可视化 [M]. 北京：首都经济贸易大学出版社，2016.

[3] CHEN C M, IBEKWE-SANJUAN F, HOU J H. The structure and dynamics of co-citation clusters: a multiple-perspective co-citation analysis[J]. Journal of the American Society for Information Science and Technology, 2010, 61(7): 1386-1409.

[4] 侯剑华，杨秀财，周莉娟. 国际图书情报领域研究的前沿主题及其演化趋势分析 [J]. 图书情报工作，2016, 60(13) : 82-90.

[5] 司红运，施建刚，陈进道，等. 从《中国人口·资源与环境》审视国内的可持续发展研究——主题脉络、知识演进与新兴热点 [J]. 中国人口·资源与环境，2019, 29(7): 166-176.

[6] 科学技术报告、学位论文和学术论文的编写格式：GB 7713—87[S]. 北京：全国文献工作标准化技术委员会，1987.

[7] 陈必坤，詹长静. 国家基金项目视角下“图书情报与档案管理”学科结构的可视化分析 [J]. 情报杂志，2017, 36(7): 105-110.

[8] 任中杰，张鹏，李思成，等. 基于微博数据挖掘的突发事件情感态势演化分析——以天津 8·12 事故为例 [J]. 情报杂志，2019, 38(2): 140-148.

[9] 单晓红，庞世红，刘晓燕，等. 基于事理图谱的网络舆情演化路径分析——以医疗舆情为例 [J]. 情报理论与实践，2019, 42(9): 99-103, 85.

[10] 赖宝春，戴瑞卿，吴振强，等. 辣椒健康植株与患枯萎病植株根际土壤细菌群落多样性的比较研究 [J]. 福建农业学报，2019, 34(9): 1073-1080.

[11] 刘政，李颖，朱培，等. 浙江省长兴县湿地维管植物多样性及区系 [J]. 浙江农林大学学报，2020,37(3):465-471.

[12] 周育臻，吴鹏飞. 贡嘎山东坡森林小型土壤节肢动物群落多样性与时空分布 [J]. 生态学杂志，2020, 39(2): 586-599.

[13] 纪昌品，王华. 鄱阳湖湿地植物群落分布特征及其对土壤环境因子的响应 [J]. 生态环境学报，2018, 27(8): 1424-1431.

[14] OLAWUMI T O, CHAN D W M. A scientometric review of global research on sustainability and sustainable development[J]. Journal of cleaner production, 2018, 183: 231-250.

[15] 学术出版规范 关键词编写规则：CY/T 173—2019[S]. 北京：国家新闻出版署，2019.

[16] 黄晨，赵星，卞杨奕，等. 测量学术贡献的关键词分析法探析 [J]. 中国图书馆学报，2019, 45(6): 84-99.

作者贡献说明：

李继红：确定论文思路，分析数据，撰写论文；

江珊：收集数据；

徐桂珍：提出修改意见；

王洪江：提出修改意见。

ChinaXiv合作期刊
知识管理论坛，2021（1）:46-55
DOI: 10.13266/j.issn.2095-5472.2021.005

Diversity of Keywords and Title Segmentation Words and a Comparison of Their Knowledge Mappings

Li Jihong Xu Guizhen Jiang Shan Wang Hongjiang
Institute of Agricultural Economy and Information, Anhui Academy of Agricultural Sciences, Hefei 230031

Abstract: [Purpose/significance] This paper aims to explore the diversity of keywords and title segmentation words, the differences of the knowledge mappings drawn based on them. **[Method/process]** We selected papers related to "academic misconduct" from CNKI from 2010 to 2019, used diversity indexes to analyze the characteristics of keywords and title segmentation words quantitatively, and compared their knowledge mappings by Citespace software qualitatively. **[Result/conclusion]** The results have shown that the richness index (S), Shannon-Wiener diversity index (H') and Pielou evenness index (E_H) of keywords were different from those of title segmentation words, and the similarity of these two units was low, indicating the keywords and title segmentation words are two different units in this paper. Although there are differences in the knowledge mappings drawn based on keywords and title segmentation words, both of them can demonstrate the research topics in the field of "academic misconduct" from their perspectives.

Keywords: academic misconduct keyword title Chinese word segmentation diversity knowledge mapping